一种基于遗传算法优化的大数据特征选择方法 *

张文杰 [1,2]，蒋烈辉 [1,2]

(1. 解放军信息工程大学 网络空间安全学院，郑州 450001; 2. 数字工程与先进计算国家重点实验室，郑州 450001)

摘 要：特征选择是大数据集预处理的重要方法，能够使后续的数据分析与处理更加高效准确。提出了一种基于遗传算法的大数据特征选择算法。该算法首先对各维度的特征进行评估，根据每个特征在同类最近邻和异类最近邻上的差异度调整其权重，基于特征权重引导遗传算法的搜索，以提升算法的搜索能力和获取特征的准确性；然后结合特征权重计算特征的适应度，以适应度作为评价指标，启动遗传算法获取最优的特征子集，并最终实现高效准确的大数据特征选择。通过实验分析发现，该算法能够有效减小分类特征数，并提升特征分类准确率。

关键词：大数据；特征选择；遗传算法；特征子集

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2018.05.0495

Using genetic algorithm for feature selection optimization on big data processing

Zhang Wenjie[1, 2], Jiang Liehui[1, 2]

(1. *Faculty of Cyberspace Security*, *PLA Information Engineering University*, *Zhengzhou* 450001, *China*; 2. *State Key Laboratory of Mathematical Engineering & Advanced Computing*, *Zhengzhou* 450001, *China*)

Abstract: Feature selection is an important method of big data set preprocessing, can make subsequent data analysis and processing more efficient and accurate. This paper proposes a novel feature selection method based on genetic algorithm for big data processing. Firstly, our method evaluates the features of each dimension, adjusts its weight according to the difference of each feature on the similar nearest neighbor and the heterogeneous nearest neighbor, and guides the search of genetic algorithm based on the feature weight, thus improves the search ability of the algorithm and the accuracy of feature acquisition. And then combines the feature weights to calculate the fitness of the feature, takes fitness as the evaluation index, and starts the genetic algorithm to obtain the optimal feature subset, finally achieve an efficient and accurate big data feature selection. The results of experiment show that our method can effectively reduce the number of classification features and improve the accuracy of feature classification.

Key words: big data; feature selection; genetic algorithm; feature subset

0 引言

随着互联网通信、数据存储、信息处理等技术的快速发展，大数据量和数据维数与日俱增。大数据集中往往包含上千个特征维度。海量的数据和特征维度包含着大量的冗余数据和无效特征，严重影响和限制了大数据分析与挖掘的性能[1,2]。为解决上述问题，特征选择通过从大数据集中剔除冗余信息，提取出具有代表性的特征子集，从而实现对大数据规模和维度的精简，提升大数据分析和处理的效率。近年来，随着大数据分析与处理技术逐步扩展深入，特征选择方法开始受到研究者的广泛关注，特征选择技术也被广泛应用于大数据聚类、文本分类、多媒体分析等诸多场景[3,4]。

特征选择的核心是通过数据处理方法提取代表性的特征子集。特征选择方法的作用主要有三项：a）通过选择大数据集合中的部分特征数据，大大减少需要分析和处理数据规模，降低大数据后期分析和处理的计算量、复杂度；b）特征选择删除大量不相关或冗余的数据信息，使得大数据易于理解和解释，更便于后期的数据处理；c）特征选择能够有效降低大数据集的维度，能够克服海量维度对大数据挖掘的限制，从而提升机器学习等方法的准确性和有效性。特征选择能够降低存储空间，计算开销等，还能揭示大数据集中隐藏的潜在结构模式和规律，对于后期的大数据挖掘和分析具有重要促进作用。

当前，特征选择方法主要包括包装法、嵌入法和过滤法三种[5]。嵌入法融合了过滤法与包装法，能够大大缩减计算时间；但嵌入法集中在局部空间内搜索，覆盖范围有限。文献[6]提出了一种改进多目标人工蜂群算法的特征选择方法，将大数据特

收稿日期：2018-05-06；**修回日期：**2018-06-28 **基金项目：**河南省基础前沿课题（142300410090）；河南省科技攻关计划项目（162102210035）

作者简介：张文杰（1978-），男，河南郑州人，工程师，硕士，主要研究方向为大数据技术及其应用（zhaifei651@163.com）；蒋烈辉（1967-），男，河南郑州人，教授，博士，主要研究方向为计算机体系结构、大数据技术及其应用.

征选择问题转换为多目标优化问题，从而提升特征选择的效率。文献[7]提出了一种新的特征子集区分度衡量准则。区别于以往仅仅考虑单个特征对区分度指标的影响，新准则将所有特征同时纳入综合考虑，计算整体特征对区分度指标的总体影响，并结合以支持向量机作为为分类工具，引导特征选择过程。文献[8]基于人工蜂群算法，提出了一种改进的特征选择优化算法，在减少特征数量和计算量的同时，以提升特征选择的效率和准确性。文献[9]提出一种基于多准则融合的特征选择算法，区别于传统的单一准则量化的思路，同时引入多种准则选取数据特征，提升特征子集多样性和算法搜索能力。然而特征选择方法多集中于考虑单个特征的重要性，使得特征重要性考量往往过于简化，忽略了不同特征之间的关联性，以及关联性对特征重要度的影响，进而降低了大数据特征选择的整体性能。

为了实现高效的特征选择，本文提出了一种基于遗传算法的启发式特征选择算法。该算法首先对各维度的特征进行评估，根据每个特征在同类最近邻和异类最近邻上的差异度调整其权重；然后结合特征权重计算特征的适应度，以适应度作为评价指标，启动遗传算法获取最优的特征子集，并最终实现高效准确的大数据特征选择。特征选择算法能够显著降低大数据分析、处理的计算时间，并提升大数据挖掘、数据分析的精确度和有效性。

1　研究背景概述

特征选择是指从完整的大数据集 D 中，基于相应策略选择出一个 k （$k<M$）维的特征子集 F，并将该特征子集应用于后期的数据分析、处理过程中。在大数据特征选择过程中，一般认为有两类属性在大数据集中并不必要：a）与目标数据不相关的属性；b）相对于目标数据而言，存在冗余属性。为了把这两类不必要的属性减到最少，需要通过特征选择对大数据集进行精简。特征选择是从大数据集合中选择属性子集的过程，通过辨别重要的属性，去除不相关或不需要的属性冗余，获取精简提炼后的大数据。特征选择在数据挖掘、机器学习等领域都有着广泛而深入的应用，是大数据分析和处理领域里非常重要的预处理方法。

近年来，有研究发现[10,11]遗传算法（genetic algorithm, GA）非常适合应用于大数据的特征选择问题。遗传算法是一种随机搜索方法，该算法能够直接对处理的对象进行操作，不受复杂的可导性、可微性或连续性等条件的限制，且在迭代过程中不受固有规则限制，可依据选择概率自主调整搜索方向，具有广泛的适应性和强大的全局搜索能力。在大数据环境中，当数据空间维度较高而对数据的内部特征无从了解时，遗传算法可以通过启发式地自学习获取优化的特征提取结果。

现有大部分特征选择算法大多采用单个评价准则，未充分考虑同类特征与异类特征的权重的差别，从而无法有效地引导遗传算子的搜索过程，使得大数据特征选择的遗传算子变异缺乏目的性，限制了算法的整体性能。基于此，本文提出了一种基于遗传算法的启发式特征选择算法。该算法首先计算每个特征在同类最近邻和异类最近邻上的差异度，综合调整其权重；然后结合特征权重计算特征的适应度，以特征适应度引导遗传算法的变异和搜索，以提升特征选择算法的搜索性能，并最终实现高效准确的大数据特征选择。

2　基于遗传算法的大数据特征选择算法

2.1　算法架构

特征选择是大数据预处理一个重要步骤，能够有效删除大数据中的冗余属性，提升大数据后期处理的效率，并能有效改善大数据分析的性能。特征选择的实质就是通过搜索迭代，从大数据中获取最具代表性的特征子集，根据评价准则评估其重要性后再进行迭代选择，直至获取最优的特征子集。

如图 1 所示，特征选择的迭代过程主要包括特征评估、子集产生和迭代停止准则三个重要步骤。由于大数据具有数据量大和特征维度高等特点，本文采用了基于遗传算法的启发式特征选择方法，首先综合每个特征的同类最近邻和异类最近邻评估其特征权重，并结合该权重计算特征适应度，以此引导遗传算法的特征搜索，提升大数据环境下特征选择的精确度。

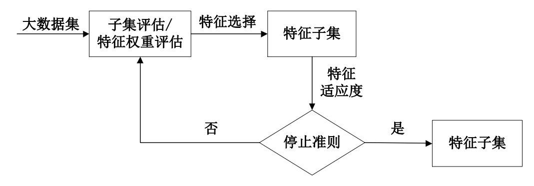


图 1　特征选择的算法框架

Fig.1　Algorithm framework for feature selection

2.2　特征权重评估

当前，有大数据研究发现[10,12]：在大数据集中，属于同一类型且距离相近的数据项都具有相似的数据特征，而距离相近但属于不同类型的数据项的数据特征差异较大。

基于此，特征权重评估算法的设计流程为：在大数据集 D 中随机选择一个数据项 x_i，在数据集中搜索其同类最近邻 $x_i(h)$ 及异类最近邻 $x_i(m)$；分别计算各维度的特征与同类最近邻的差异度值，与异类最近邻的差异度值，根据两者的特征差异度相应地调整其权重。通过反复迭代，最后选择权重值最高的 k 维特征组成新的特征子集。

$diff(x_1,x_2,j)$ 表示数据项 x_1 和 x_2 在特征 f_j 上的差异度。在数据项 x_1 与 x_2 上,两者在特征维度 f_j 上的差异度是

$$diff(x_1,x_2,j)=\frac{|x_{1j}-x_{2j}|}{\max(f_j)-\min(f_j)} \tag{1}$$

在每轮迭代过程中，根据 x_i 的 r 个同类最近邻 $x_i(h)$ 和 r 个异类最近邻 $x_i(m)$ 在特征 f_j 上的差异度，调整 x_i 关于特征 f_j 上的权值：

$$\omega(j)=\omega(j)-\sum_{h\subseteq r}diff(x_i,x_i(h),j)/M+\sum_{\delta\subseteq r}diff(x_i,x_i(\delta),j)/M \tag{2}$$

特征权重评估算法的具体描述如下：

输入：大数据集 D，迭代次数 t，子集维数 k 。

输出：特征权重 $\omega(1,\cdots,M)$ 。

初始化大数据集 D 中特征权重 $\omega(1,\cdots,M)=0.5$；
for $i=1$ to t do
随机获取一个数据项 x_i；
搜索 x_i 的 r 个同类最近邻 $x_i(h)$ 和 r 个异类最近邻 $x_i(m)$
for $j=1$ to M do

$$\omega(j)=\omega(j)-\sum_{h\subseteq r} diff(x_i,x_i(h),j)/M+\sum_{\delta\subseteq r} diff(x_i,x_i(\delta),j)/M$$

end for
更新特征权重 $\omega(1,\cdots,M)$；
end for

特征权重评估算法将每个样本数据项与其 r 个同类最近邻和 r 个异类最近邻相比，根据该样本数据项与邻域数据项在相关特征维度上的差异度调整其权值。同类差异度越小，异类差异度越大，说明该特征维度越具有代表性，权重增加；同类差异度越大，异类差异度越小，说明该特征维度越缺乏代表性，权重减小。相比一般的特征选择算法，该算法综合考虑了特征的同类、异类近邻与各维度的相关性，使得特征权重更能客观反映该维度特征的代表性，能够使后续遗传算法的特征搜索与选择的性能更优，鲁棒性更强。

2.3 基于遗传算法的特征选择方法

本文采用了基于遗传算法的启发式特征选择方法，首先综合每个特征的同类最近邻和异类最近邻评估其特征权重，并结合该权重计算特征适应度，以此引导遗传算法的特征搜索，提升大数据环境下特征选择的精确度。具体的操作步骤如下。

a)随机生成初始种群 $X=\{x_i^0,\cdots x_N^0\}$，种群规模为 N 。对解空间进行编码、初始化。

b)根据设定好的适应度函数计算第 t 代全部个体的适应度值 $f(x_i^t)$， $i=1,\cdots,N$ 。

c)综合比较特征子集的类间、类内距离，以特征子集的类间距离与类内距离之比为适应度函数。

$$f=\frac{\sum_{i=1}^{c}\|\bar{x}(i)-\bar{x}\|^2}{\sum_{i=1}^{c}\frac{1}{n_i-1}\sum_{j=1}^{n_i}\|\overline{x_j}(i)-\bar{x}(i)\|^2} \tag{3}$$

其中：$\bar{x}$ 表示特征子集在大数据集的均值向量；$\bar{x}(i)$ 表示特征子集在 i 类的均值向量；$\overline{x_j}(i)$ 表示第 i 类的第 j 个样本向量；n_i 为第 i 类的样本个数；c 为类别数。类间距离越大，类内距离越小，说明该特征子集的适应度越高；反之，类间距离越小，类内距离越大，说明该特征子集的适应度越低。

d)综合考虑特征权重评估算法获取的特征权重与其适应度值，估算特征选取概率，从上一代种群 $x_1^t,\cdots,x_N^t$ 中下一轮迭代的种群 $x_1^{t+1},\cdots,x_N^{t+1}$。其中 x_i^t 被选择的概率为

$$p(x_i^t)=\frac{\omega(x_i)f(x_i^t)}{\sum_{i=1}^{N}\omega(x_i)f(x_i^t)},i=1,\cdots,N \tag{4}$$

e)以相同的概率从更新的种群 $x_1^{t+1},\cdots,x_N^{t+1}$ 中选择两个个体，以概率 P_c 完成染色体的交叉重组。同时，以概率 P_m 对个体的某基因位进行突变，获取新一代的种群 $x_1^{t+2},\cdots,x_N^{t+2}$ 。获取 $x_1^{t+2},\cdots,x_N^{t+2}$ 中适应度值最高的个体 x_*^{t+2} 。

f)比较 x_*^{t+2} 的适应度值，若大于等于相关的适应度阈值，或迭代次数已经达到最大，则终止过程，输出 x_*^{t+2}；否则，令 $t=t+1$，跳转进入步骤 b)。

g)从大数据 D 选择适应度的排名前 k 个特征，组成特征子集 F 。

3　实验结果及分析

实验数据主要基于业界标准的 UCI 数据库，从中选取比较具有代表性的 10 个数据集作为测试数据。数据集的具体信息见表 1 。在本文的 UCI 数据集[13]实验中，k 值设置为 10。每次测试轮流选择 1 个数据集作为独立的测试数据集，其余 9 个数据集作为训练模型。该测试数据集包含 10 个多分类的数据集，样本数为 150~569，特征数规模为 4~255。这 10 个数据集数据类型各不相同，数据特征具有广泛的代表性，能够全面有效地衡量和比较各种算法特征选择的性能指标。实验环境为 Lenovo M9620T 的台式电脑，Intel(R) 6 Core(TM) i3-3240 3.39 GHz CPU，4.0 GB 内存，Windows7 64 位操作系统，软件环境为 MATLAB R2010b。

为全面比较本文算法与同类特征选择算法的性能，实验将其分别与 GA_SVM 算法[14]（基于遗传算法）、ReliefF 算法[15]（传统特征选择算法）进行比较。GA_SVM 算法、ReliefF 算法分别是各自领域内具有代表性的算法。表 2 是本文算法与基于遗传算法方法的分类准确率比较结果。在实验中，迭代次数 t 一般依据经验或者多次实验获取，10 次重复计算以估计两种方法在 UCI 数据集上的分类准确率，得到的分类准确率以平均百分比±标准差来表示。

如表 1 所示，在 10 个数据集中，由于 Iris 数据集的特征数与类别数较少，GA_SVM 算法与本所算法的分类准确率相同；在其余 9 个数据集上，本文算法的性能都优于 GA_SVM 算法，分类准确率都有了不同幅度的提升。同时，只有在 Dermatology 数据集上，本文算法分类准确率的标准差略高于 GA_SVM 算法；其余 9 个数据集上，本文算法分类准确率的标准差都要低于 GA_SVM 算法，这说明本文算法在分类准确率、分类稳定性上都要优于 GA_SVM 算法。

表 3 是本文算法与传统特征选择算法的分类准确率比较结果。同样地，在实验中 10 次重复计算以估计两种方法在 UCI 数

据集上的分类准确率，得到的分类准确率和选择特征数以平均百分比±标准差来表示。

如表 3 所示，相比传统的 ReliefF 算法，本文算法在分类准确率、选择特征数上的性能改善更为明显。在 10 个数据集中，相比 ReliefF 算法，本文算法的分类准确率都有了不同幅度的提升，选择的特征数均值也小于同类算法。同时，在 Iris 数据集上，本文算法分类准确率的标准差略高于 GA_SVM 算法，其余数据集上的标准差都低于 ReliefF 算法；在 Dermatology 数据集上，本文算法选择的特征数标准差略高于 GA_SVM 算法，其余数据集上的标准差都低于 ReliefF 算法，这说明本文算法在分类准确率、分类稳定性上都要优于 ReliefF 算法，具有最高的分类准确率、最少的选择的特征数。

表 1　选择测试的 UCI 数据集

Table 1　Select UCI data sets for testing

数据集	样本个数	特征数	类别数
Handwrite	323	255	2
WPBC	194	33	2
Thyroid-disease	215	5	3
Glass	214	9	2
WDBC	569	30	2
Ionosphere	351	34	2
Iris	150	4	3
Wine	178	13	3
Heart disease	297	13	5
Dermatology	358	34	6

表 2　本文方法和基于遗传算法方法的分类准确率比较/%

Table 2　Comparison of classification accuracy between this method and genetic algorithm/%

数据集	本文方法	GA_SVM 算法
Iris	100.00±0.00	100.00±0.00
Dermatology	99.00±1.66	96.19±1.24
Glass	86.10±1.97	85.60±1.96
Handwrite	95.56±2.34	94.80±3.32
Ionosphere	99.43±1.21	98.56±2.03
WDBC	91.59±2.14	88.10±2.25
WPBC	83.84±5.14	81.50±7.13
Wine	99.00±2.11	98.00±3.50
Thyroid-disease	88.24±1.47	84.06±3.54
Heart disease	99.60±0.71	99.30±0.82

4　结束语

针对当前大数据维度爆炸、计算复杂度高等问题，本文提出了一种基于遗传算法的大数据特征选择算法，以消除大数据集中的冗余特征，提升特征子集的选择准确率。在 10 个 UCI 数据集上进行实验发现，相比其他的特征选择算法，本文算法能够有效地提升大数据分类的准确率，并减少特征子集的特征数量。同时，本文算法特征选择的稳定性也有一定程度的改善。综合而言，本文的特征选择算法能够高效准确地获取大数据特征子集，能够显著降低大数据后续分析、处理的计算复杂度。

表 3　本文方法和不带特征染色体遗传算法的实验结果

Table 3　Experimental results of this method and genetic algorithm without characteristic chromosom

数据集	特征数	本文方法		ReliefF 算法	
		分类准确率%	选择的特征数	分类准确率%	选择的特征数
Iris	4	100.00±0.00	1.2±0.28	96.00±3.44	1.8±0.38
Dermatology	34	99.00±1.66	13.9±3.45	98.57±2.02	15.4±3.32
Glass	9	86.10±1.97	3.7±1.26	81.97±5.34	5.1±1.63
Handwrite	255	95.56±2.34	11.2±1.71	91.74±2.32	12.3±2.72
Ionosphere	34	99.43±1.21	10.3±1.76	94.80±2.10	11.8±3.33
WDBC	30	93.59±2.14	6.2±1.12	91.11±2.58	7.0±1.05
WPBC	33	93.84±3.18	2.5±0.88	90.04±5.14	2.9±0.99
Wine	13	100.00±0.00	4.2±0.50	99.44±1.76	4.6±0.72
Thyroid-disease	5	88.24±1.47	2.8±0.99	81.43±7.29	3.2±1.14
Heart disease	13	89.60±0.71	5.2±2.15	86.81±3.64	6.7±3.16

参考文献：

[1] 周琪. 特征选择与特征学习算法研究 [D]. 合肥：中国科学技术大学, 2017. (Zhou Qi. Research on feature selection and feature learning algorithm [D]. Hefei: University of Science and Technology of China, 2017.)

[2] 张钧波. 面向大数据的高效特征选择与学习算法研究 [D]. 成都：西南交通大学, 2015. (Zhang Junbo. Research on efficient feature selection and learning algorithms for big data [D]. Chengdu: Southwest Jiaotong University, 2015.)

[3] 李俊. 基于智能优化的特征选择及分类方法研究 [D]. 武汉：武汉大学, 2014. (Li Jun. Research on feature selection and classification based on intelligent optimization algorithms [D]. Wuhan: Wuhan University, 2014.)

[4] 吕辉. 基于大数据和高维数据的聚类方法的研究与设计实现 [D]. 昆明：云南大学, 2015. (Lyu Hui. Research and design of clustering method based on big data and High-dimensional data s [D]. Kunming: Yunnan University, 2015.)

[5] 王翔，胡学钢. 高维小样本分类问题中特征选择研究综述 [J]. 计算机应用, 2017, 37 (9): 2433-2438. (Wang Xiang,　Hu Xuegang. Overview on feature selection in high-dimensional and small-sample-size classification [J]. Journal of Computer Applications, 2017, 37 (9): 2433-2438.)

[6] 巢秀琴，李炜. 基于改进多目标人工蜂群算法的特征选择方法 [J]. 计算机科学与探索, 2018, 16 (1): 71-78. (Chao Xiuqin, Li Wei. A feature selection method optimizedby artificial bee colony algorithm [J]. Journal of Frontiers of Computer Science and Technology, 2018, 16 (1): 71-78.)

[7] 谢娟英，谢维信. 基于特征子集区分度与支持向量机的特征选择算法 [J]. 计算机学报, 2014, 37 (8): 1704-1718. (Xie Juanying, Xie Weixin. Several feature selection algorithms based on the discernibility of a feature subset and SupportVector machines [J]. Chinese Journal of Computers, 2014,



37 (8): 1704-1718.)

[8] Hancer E, Xue Bing, Zhang Mengjie, *et al.* Pareto front feature selection based on artificial bee colony optimization [J]. Information Sciences, 2018, 422 (1): 462-479.

[9] 关晓颖，陈果，林桐. 特征选择的多准则融合差分遗传算法及其应用 [J]. 航空学报, 2016, 37 (11): 3455-3465. (Guan Xiaoyin, Chen Guo, Lin Tong. Feature selectionmethod based on differential evolution and genetic algorithm with multi-criteria evaluation and its applications [J]. Acta Aeronautica et Astronauticas Sinica, 2016, 37 (11): 3455-3465.)

[10] 赵荣珍，李坤杰. 基于类内类间判据与遗传算法的故障特征选择方法 [J]. 兰州理工大学学报, 2017, 43 (2): 35-39. (Zhao Rongzhen; Li Kunjie. Fault feature selection method based on within-class and among-class criterion and genetic algorithm [J]. Journal of Lanzhou University of Technology, 2017, 43 (2): 35-39.)

[11] 王娜. 基于遗传算法的混合特征选择方法研究 [D]. 西安：陕西师范大学, 2015. (Wang Na. Study on hybrid feature selection method based on genetic algorithm [D]. Xi'an: Shaanxi Normal University, 2015.)

[12] 陈磊. 文本表示模型和特征选择算法研究 [D]. 合肥：中国科学技术大学，2017. (Chen Lei. Text representation model and feature selection algorithm [D]. Hefei: University of Science and Technology of China, 2017.)

[13] Murphy P M, Aha D W. UCI repository of machine learning database [DB/OL]. (2006-05-12) . http://www. ics. uci. edu/mlearn/MLRepository. html.

[14] Huang C, Wang C. A GA-based feature selection and parameters optimization for support vector machines [J]. Expert Systems with Applications, 2016, 31 (2): 231-240.

[15] 伍杰华. 基于 RReliefF 特征选择算法的复杂网络链接分类 [J]. 计算机工程，2017, 43 (8): 208-214. (Wu Jiehua. Complex network link classification based on RReliefF feature selection algorithm [J]. Computer Engineering, 2017, 43 (8): 208-214.)